CGU: An Algorithm for Molecular Structure Prediction

K.A. Dill, A.T. Phillips, J.B. Rosen

Abstract. A global optimization method is presented for predicting the minimum energy structure of small protein-like molecules. This method begins by collecting a large number of molecular conformations, each obtained by finding a local minimum of a potential energy function from a random starting point. The information from these conformers is then used to form a convex quadratic global underestimating function for the potential energy of all known conformers. This underestimator is an L_1 approximation to all known local minima, and is obtained by a linear programming formulation and solution. The minimum of this underestimator is used to predict the global minimum for the function, allowing a localized conformer search to be performed based on the predicted minimum. The new set of conformers generated by the localized search serves as the basis for another quadratic underestimation step in an iterative algorithm. This algorithm has been used to predict the minimum energy structures of heteropolymers with as many as 48 residues, and can be applied to a variety of molecular models. The results obtained also show the dependence of the native conformation on the sequence of hydrophobic and polar residues.

1. Introduction

It is widely accepted that the folded state of a protein is completely dependent on the one-dimensional linear sequence (i.e. “primary” sequence) of amino acids from which it is constructed: external factors, such as helper (chaperone) proteins, present at the time of folding have no effect on the final, or native, state of the protein. Furthermore, the existence of a unique native conformation, in which residues distant in sequence but close in proximity exhibit a densely packed hydrophobic core, suggests that this 3-dimensional structure is largely encoded within the sequential arrangement of these hydrophobic (H) and polar (P) amino acids. The assumption that hydrophobic interaction is the single most dominant force in the correct folding of a protein also suggests that simplified potential energy functions, for which the terms involve only pairwise H-H attraction and steric overlap repulsion, may be sufficient to guide computational search strategies to the global minimum representing the native state.

During the past 20 years, a number of computer algorithms have been developed that aim to predict the fold of a protein (see for example [3], [5], [8], [10]). Such approaches

1991 Mathematics Subject Classification. Primary 65K05; Secondary 90C26, 90C90.

The first author was supported by the NSF grant BIR-9119575; the second author was supported by NIH grant RR-08605 and the San Diego Supercomputer Center; the third author was supported by the ARPA/NIST grant 60NANB2d1272 and NSF grant CCR-9509085.

are generally based on two assumptions. First, that there *exists* a potential energy function for the protein; and second that the folded state corresponds to the structure with the lowest potential energy (minimum of the potential energy function) and is thus in a state of thermodynamic equilibrium. This view is supported by in vitro observations that proteins can successfully refold from a variety of denatured states.

2. A Simple Polypeptide Model

Computational search methods are not yet fast enough to find global optima in real-space representations using accurate all-atom models and potential functions. A practical conformational search strategy will require both a simplified molecular model with an associated potential energy function which consists of the dominant forces involved in protein folding, and also a global optimization method which takes full advantage of any special properties of this kind of energy function. In what follows, we describe a global optimization algorithm which has been successfully used for one such simplified model. We then describe a more realistic model, which we believe will permit the use of our algorithm on small protein molecules.

In our initial testing of the CGU algorithm (to be described shortly), we chose to use a simple "string of beads" model consisting of *n* monomers C connected in sequence (see Figure 2.1). The 3-dimensional position of each monomer, relative to the previous monomers in the sequence, is defined by the parameters *l* (the "bond lengths"), θ (the "bond angles"), and ξ (the backbone "dihedral angles"). Of these we have chosen to fix *l* and θ (the reasons for this will become clear later), thus reducing the number of independent parameters necessary to uniquely define a 3-dimensional conformation to only *n*-1. In order to model the H-P effects that are encoded within the backbone sequence, each "bead" C in this simplified model is categorized as either hydrophobic (H) or polar (P).

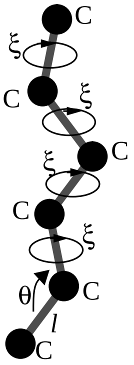

Figure 2.1 Simplified "String of Beads" Polypeptide Model

Corresponding to this simplified polypeptide model is a potential energy function also characterized by its simplicity. This function includes just three components: a contact energy term favoring pairwise H-H residues, a steric repulsive term which rejects any conformation that would permit unreasonably small interatomic distances, and a main chain torsional term that allows only certain preset values for the backbone dihedral angles ξ. Despite its simplicity, the use of this type of potential function has already proven successful in studies conducted independently by Sun, Thomas, and Dill [13] and by Srinivasan and Rose [11]. Both groups have demonstrated that this type of potential function is sufficient to accurately model the forces which are most responsible for folding proteins. Here our energy function is somewhat different from either of those. The specific potential function used initially in this study has the following form:

$$E_{total} = E_{ex} + E_{hp} + E_{\xi} \tag{1}$$

where the steric repulsion and hydrophobic attraction terms $E_{ex} + E_{hp}$ can conveniently be combined and represented by the well known Lennard-Jones pairwise potential function

$$\sum_{|i-j|>2} \varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 2H_{ij}\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) \ .$$

This term defines the potential energy contributions of all beads separated by more than two along the primary chain. The Lennard-Jones coefficients ε_{ij} and σ_{ij} are constants defined by the relationships between the two specific beads (e.g. amino acids) involved. The terms involving r_{ij} in the Lennard-Jones expression represent the Euclidean distances between beads i and j. The constant $H_{ij} = 1$ if beads i and j are both H-type (attractive monomers), and hence both a repulsive force (ensuring that the chain is “self-avoiding”) and an attractive force (since the beads are H-H) are added to the potential energy (see Figure 2.2). On the other hand, $H_{ij} = 0$ if the beads i and j are H-P, P-H, or P-P pairs, so that the Lennard-Jones contribution to the total potential energy is just the repulsive force that ensures self-avoidance.

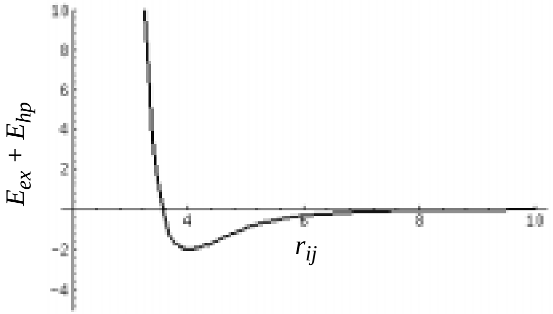

Figure 2.2 Lennard-Jones Pair Potential with $H_{ij} = 1$, $\varepsilon_{ij} = 2$, and $\sigma_{ij} = 4$

A trigonometric based penalty implementing the potential energy term E_{ξ} in equation 1 was used in these tests, and had the following form:

$$E_{\xi} = \sum_{i} C_1(1 + \cos(3\xi_i)) \ .$$

Using this term, there are only three “preferred” backbone dihedral angles of 60˚, 180˚, and 300˚ with all others penalized to some extent (determined by the constant C_1). The purpose of this term is to mimic, in some elementary sense, the restrictive nature of the Ramachandran plot (see [4]) for each residue in a realistic protein model.

3. The CGU Global Optimization Algorithm

One practical means for finding the global minimum of the polypeptide’s potential energy function is to use a global underestimator to localize the search in the region of the global minimum. This CGU (convex global underestimator) method is designed to fit all known local minima with a convex function which underestimates all of them, but which differs from them by the minimum possible amount in the discrete L_1 norm (see Figure 3.1). Any non-negative linear combination of convex functions can be used for the underestimator, but for simplicity we use convex quadratic functions. The minimum of this underestimator is used to predict the global minimum for the function, allowing a localized conformer search to be performed based on the predicted minimum. A new set of conformers generated by the localized search then serves as a basis for another quadratic underestimation. After several repetitions, the global minimum can be found with reasonable assurance.

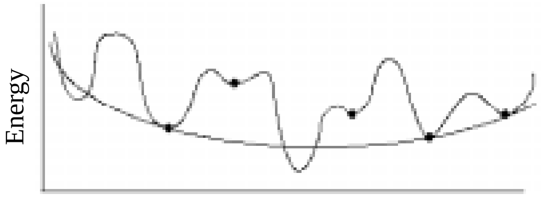

Figure 3.1 The Convex Global Underestimator (CGU)

This method, first described in [9], is presented in terms of the differentiable potential energy function $E_{total}(\phi)$, where $\phi \in \mathbf{R}^{n-1}$ (n represents the number of residues in the polypeptide chain), and where $E_{total}(\phi)$ has many local minima. Thus, ϕ is a vector of n-1 backbone dihedral angles. Defining $\tau = n-1$, then to begin the iterative process, a set of $k \geq 2\tau+1$ distinct local minima are computed. This can be done with relative ease by using an efficient unconstrained minimizer, starting with a large enough set of points chosen at random in an initial hyperrectangle $H\phi$, which is assumed to enclose the entire torsion angle space.

Assuming that $k \geq 2\tau+1$ local minima $\phi^{(j)}$, for j=1,...,k, have been computed, a convex quadratic underestimator function $F(\phi)$ is now fitted to these local minima so that it underestimates all the local minima, and normally interpolates $E_{total}(\phi^{(j)})$ at $2\tau+1$ points (see Figure 3.1). This is accomplished by determining the coefficients in the function $F(\phi)$ so that

$$(2) \qquad \delta_j = E_{total}(\phi^{(j)}) - F(\phi^{(j)}) \geq 0$$

for j=1,...,k, and where $\sum_{j=1}^{k} \delta_j$ is minimized. That is, the difference between $F(\phi)$ and $E_{total}(\phi)$ is minimized in the discrete L_1 norm over the set of k local minima $\phi^{(j)}$, j=1,...,k. Although many choices are possible, the underestimating function $F(\phi)$ selected for the CGU method is a separable convex quadratic given by

$$(3) \qquad F(\phi) = c_0 + \sum_{i=1}^{\tau} \left(c_i \phi_i + \frac{1}{2} d_i \phi_i^2 \right).$$

Note that c_i and d_i appear linearly in the constraints of equation 2 for each local minimum $\phi^{(j)}$. Convexity of this quadratic function is guaranteed by requiring that $d_i \geq 0$ for i=1,...,τ.

Additionally, in order to guarantee that $F(\phi)$ attains its global minimum F_{min} in the hyperrectangle $H\phi = \{\phi_i : 0 \leq \underline{\phi}_i \leq \phi_i \leq \overline{\phi}_i \leq 2\pi\}$, the following additional set of constraints are imposed on the coefficients of $F(\phi)$:

(4) $c_i + \underline{\phi}_i d_i \le 0$ and $c_i + \bar{\phi}_i d_i \ge 0$ for i=1,...,τ.

Note that the satisfaction of equation 4 implies that $c_i \le 0$ and $d_i \ge 0$ for i=1,...,τ.

The unknown coefficients c_i, i=0,...,τ, and d_i, i=1,...,τ, can be determined by a simple linear programming formulation and solution, and since the convex quadratic function $F(\phi)$ gives a global approximation to the local minima of $E_{total}(\phi)$, then its easily computed global minimum function value F_{min} is a good candidate for an approximation to the global minimum of the potential energy function $E_{total}(\phi)$. The complete details of this linear programming formulation are given in [9], and so are not presented here.

The convex quadratic underestimating function $F(\phi)$ determined by the values $c \in \mathbf{R}^{\tau+1}$ and $d \in \mathbf{R}^{\tau}$ provide a global approximation to the local minima of $E_{total}(\phi)$, and its easily computed global minimum point F_{min} is given by $(\phi_{min})_i = -c_i/d_i$, i=1,...,τ, with corresponding function value F_{min} given by $F_{min} = c_0 - \sum_{i=1}^{\tau} c_i^2/(2d_i)$. The value F_{min} is a good candidate for an approximation to the global minimum of the potential energy function $E_{total}(\phi)$, and so ϕ_{min} can be used as an initial starting point around which additional configurations (i.e. local minima) should be generated. These local minima are added to the set of all known local minima, and the process is repeated. Before each iteration of this process, it is necessary to reduce the volume of the hyperrectangle Hϕ over which the new configurations are produced so that a tighter fit of $F(\phi)$ to the local minima "near" ϕ_{min} is constructed.

If E_c is a cutoff energy, then one means for modifying the size of the hyperrectangle Hϕ at any step is to let Hϕ = {ϕ: $F(\phi) \le E_c$}. Clearly, if E_c is reduced, the size of Hϕ is also reduced. At every iteration the predicted global minimum value F_{min} satisfies $F_{min} \le E_{total}(\phi^*)$, where ϕ* is the smallest *known* local minimum conformation computed so far (see Figure 3.2). Therefore, $E_c = E_{total}(\phi^*)$ is often a good choice. If at least one improved point ϕ, with $E_{total}(\phi) < E_{total}(\phi^*)$, is obtained in each iteration, then the search domain Hϕ will strictly decrease at each iteration, and may decrease substantially in some iterations (see Figure 3.3). Such a means for reducing the search domain Hϕ does not of course guarantee that the true global minimum will be found. In fact, it should be clear that the

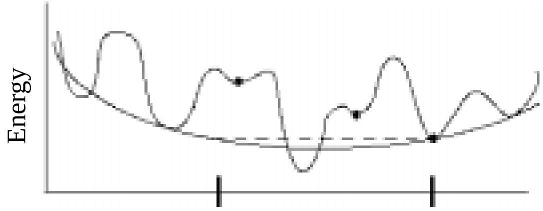

Figure 3.2 Defining the Hyperrectangle Hϕ

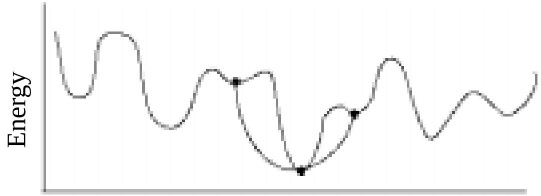

Figure 3.3 The New CGU Over a Reduced Hyperrectangle Hϕ

true global solution may be excluded from the new search domain if it avoids detection as a local minimum solution at every iteration. Hence it is very important that the initial set of k distinct local minima be sufficiently large so that either the true global minimum is included among them, or so that the global convex underestimator $F(\phi)$ accurately models and predicts the global structure of $E_{total}(\phi)$. As a general rule of thumb (based only on computational experience), $k = 10(2\tau+1)$ is sufficient for this purpose.

Based on the general discussion above and the details provided in [9], the CGU algorithm can be succinctly described as follows:

1. Compute $k \geq 2\tau+1$ distinct local minima $\phi^{(j)}$, for $j=1,...,k$, of the function $E_{total}(\phi)$.
2. Compute the convex quadratic underestimator function

$$F(\phi) = c_0 + \sum_{i=1}^{\tau}\left(c_i\phi_i + \frac{1}{2}d_i\phi_i^2\right)$$

   by solving a linear program (see [9] for details). The optimal solution to this linear program directly provides the values of c and d.
3. Compute the predicted global minimum point ϕ_{min} given by $(\phi_{min})_i = -c_i/d_i$, $i=1,...,\tau$, with corresponding function value F_{min} given by $F_{min} = c_0 - \sum_{i=1}^{\tau} c_i^2/(2d_i)$.
4. If $\phi_{min} = \phi^*$, where $\phi^* = \mathrm{argmin}\{E_{total}(\phi^{(j)}), j=1,2,...\}$ is the best local minimum found so far, then stop and report ϕ^* as the approximate global minimum conformation.
5. Reduce the volume of the hyperrectangle $H\phi$ over which the new configurations will be produced by using the rule $H\phi = \{\phi\colon F(\phi) \leq E_c\}$ where $E_c = E_{total}(\phi^*)$.
6. Use ϕ_{min} as an initial starting point around which additional local minima $\phi^{(j)}$ of $E_{total}(\phi)$ (restricted to the hyperrectangle $H\phi$) are generated.
7. Return to step 2.

While the number of new local minima to be generated in step 6 is unspecified, a value exceeding $2\tau+1$ would of course be required for the construction of another convex quadratic underestimator in the next iteration (step 2). In the computational tests presented in the next section, we have chosen to use $10(2\tau+1)$ starting points for both steps 1 and 6 in an attempt to generate at least $2\tau+1$ distinct local minima.

The rate and method by which the hyperrectangle size is modified are important in determining the convergence properties of the CGU algorithm. It can be shown that if the convex underestimator $F(\phi)$ does in fact underestimate the global minimum of $E_{total}(\phi)$ at every iteration of the CGU algorithm, then by appropriately applying “branch-and-bound” techniques to this method, finite convergence to the global minimum can be guaranteed (albeit in a possibly exponential number of steps). Note that the underestimator need not actually underestimate *all* local minima for this property to be true, it only need underestimate the global solution at each step. Since $F(\phi)$ underestimates all known local minima in the current hyperrectangle $H\phi$ (by construction), it is very likely that it will also underestimate the global minimum. While this is not guaranteed to be the case, our computational experience shows that is usually satisfied. Another method of constructing a convex global underestimator, and a related global optimization algorithm, is described elsewhere in [6], but no computational comparison of these two methods has yet been made.

4. Computational Results for the Simplified Model

The computational results presented below for the simplified model were obtained on a network of eight Sun SparcStations using the MPI message passing system for communication between the CPUs. Steps 1 and 6 of the algorithm (presented in section 3) were performed in parallel on all eight of the processors, while the remaining steps of the algorithm were done sequentially on a single designated “master” processor.

A detailed set of computational results for the CGU method have previously been presented in [9]. In that paper, the method was tested on a large sample of homopolymer sequences (that is, all residues are hydrophobic) ranging in size from only 4 residues (“beads”) to as many as 30. This paper presents only two additional test cases, but these tests serve to demonstrate that the CGU method can be successfully applied to larger heteropolymer sequences (i.e. mixed sequences of H and P). The two HP sequences tested were designed by E. Shakhnovich as part of a friendly competition between his group at Harvard and the Dill group at the University of California, San Francisco. In that competition, the Harvard group designed a set of 3-dimensional lattice-based 48-mer HP sequences with a known folded target structure (also restricted to the lattice) which they denoted “putative native state” (PNS). The PNS was not known to be the global solution, since it was computed by an inverse folding technique using a Monte Carlo method. The object of that competition was to see if the Dill group could find the PNS (or a folded state with an even lower energy) using their own algorithms, but given only the primary HP sequence. Ten HP sequences were tested, and of these we have selected two representative ones, the sequences labeled #8 and #10 (see [14]).

The computational tests presented below serve to illustrate two points: (1) that the CGU method can be practical for moderate size sequences (in this case 48-mer sequences), (2) and that the global minimum energy is in fact *very highly dependent* on the primary sequence.

When applied to the first of these sequences, sequence #8, the CGU method found the folded state, with an associated minimum energy of -87.57, as shown in Figure 4.1 (the dark grey beads are P type while the light grey beads are H type). For this folded conformation (which we will denote F#8), if the sequence had consisted of all P type monomers,

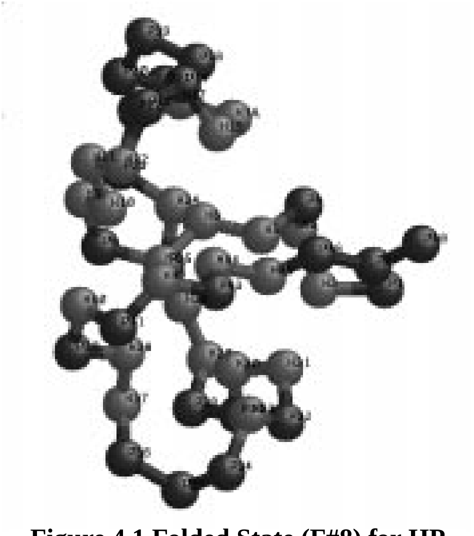

Figure 4.1 Folded State (F#8) for HP Sequence #8

Sequence #8 is PHHPHHHPHHHHPPHHHPPPPPPHPHHPPHHPHPPPHHPHPHPHHPPP then corresponding energy for F#8 would be +128.99, whereas if it had consisted of all H type monomers, the energy would have been -239.97. Furthermore, when the sequence is fully extended, i.e. all backbone dihedrals ξ are set to 180˚, then the corresponding energy for sequence #8 is -4.22. This may also be considered an upper bound.

Recall that F#8 is the presumed global minimum conformation for the HP sequence #8. If this sequence is replaced by all H or all P type monomers, then F#8 is not necessarily even a *local* minimum. Figures 4.2 and 4.3 show the conformations which result from a single local minimization (i.e. relaxation) beginning from state F#8 with these two homopolymer sequences in place of sequence #8. Clearly, they are decidedly different. Table 4.1 summarizes these various results.

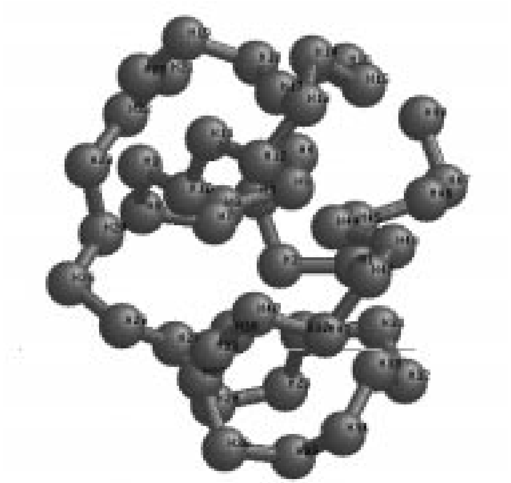

Figure 4.2 Relaxation From F#8 Using 48-mer All H-type Homopolymer

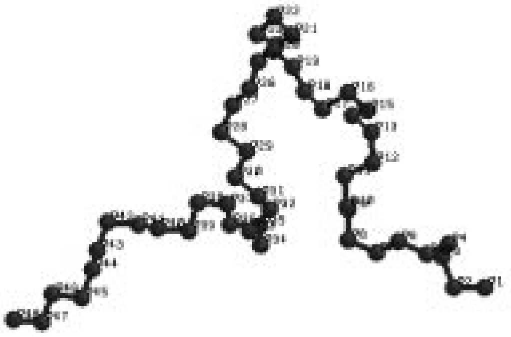

Figure 4.3 Relaxation From F#8 Using 48-mer All P-type Homopolymer

A similar analysis was performed for sequence #10. Figure 4.4 shows the folded conformation (denoted F#10) for sequence #10, which has a corresponding energy value of -97.22. Figures 4.5 and 4.6 show the relaxed conformations obtained when sequence #10 is

Sequence #10 is PHHPPPPPPHHPPPHHHPHPPHPHHPPHPPHPPHHPPHHHHHHHPPHH

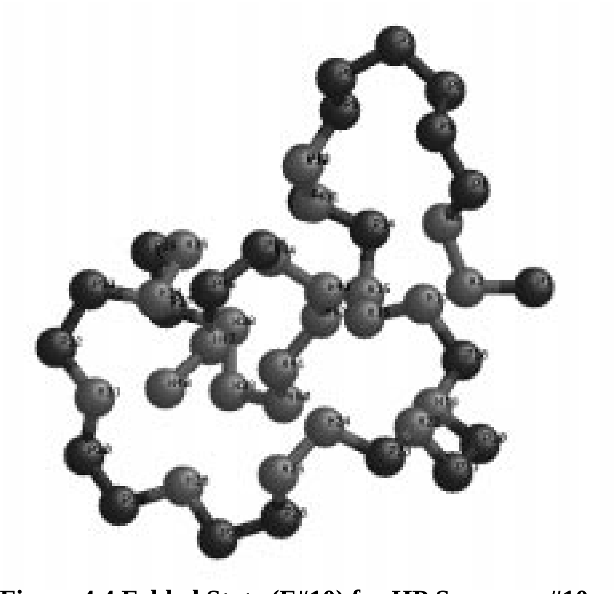

Figure 4.4 Folded State (F#10) for HP Sequence #10

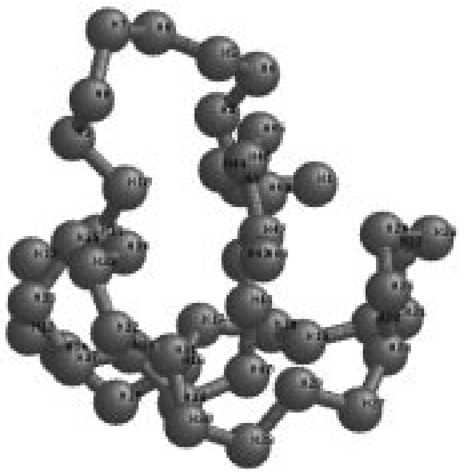

Figure 4.5 Relaxation From F#10 Using 48-mer All H-type Homopolymer

replaced by an all H and an all P homopolymer sequence, and Table 4.2 summarizes the results according to energies for each state. Like sequence #8, sequence #10 consists of 50% H and 50% P type monomers, yet the folded conformations F#8 and F#10 are decidedly different (compare Figures 4.1 and 4.4). In fact, Table 4.3 shows that the energy obtained by sequence #8 when placed into state F#10 (the "native" state for sequence #10) is actually considerably above its minimum energy in F#8 (compare -23.20 to -87.57). Likewise, when sequence #10 is evaluated in state F#8 it also obtains a much higher energy (compare +39.48 to -97.22). Furthermore, even if these "non-native" conformations are allowed to relax to a local minimum, Table 4.3 clearly shows that the result

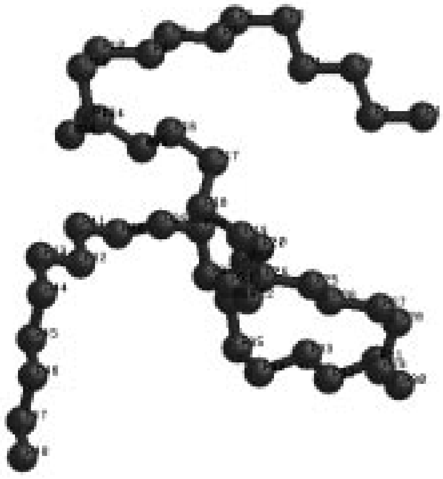

Figure 4.6 Relaxation From F#10 Using 48-mer All P-type Homopolymer

Table 4.1 Dependence of Energy on Primary Sequence (Based on #8)

	All H-type Homopolymer	HP Sequence #8	All P-type Homopolymer
Fully Extended	-19.55	-4.22	+1.27
F#8	-239.97	-87.57	+128.99
Relaxation from F#8	-334.22	-87.57	+13.06

Table 4.2 Dependence of Energy on Primary Sequence (Based on #10)

	All H-type Homopolymer	HP Sequence #10	All P-type Homopolymer
Fully Extended	-19.55	-3.98	+1.27
F#10	-243.87	-97.22	+133.91
Relaxation From F#10	-315.41	-97.22	+10.56

remains non-native. Hence, the global minimum energy and associated conformation of an HP sequence is *very highly dependent* on the primary sequence used. Figures 4.7 and 4.8 show the two “relaxed” conformations obtained by locally minimizing each sequence from the “other” sequences native state.

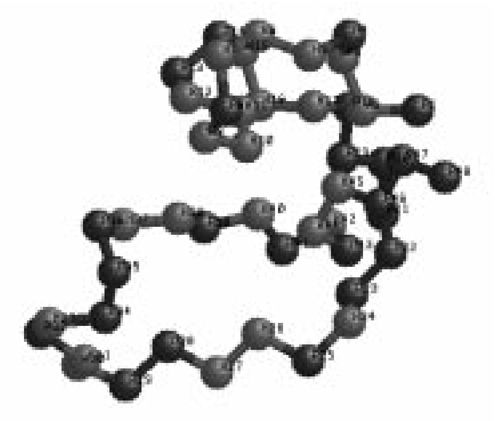

Figure 4.7 Conformation Obtained by Relaxation of Sequence #8 from State F#10

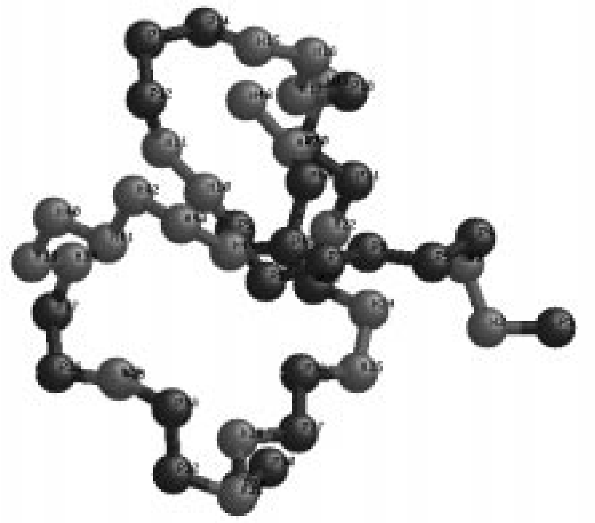

Figure 4.8 Conformation Obtained by Relaxation of Sequence #10 from State F#8

It is also the case that removing the lattice restriction, as we have done, gives a very different native conformation with the identical HP sequence. This is seen by comparing the conformations in Figures 4.1 and 4.4 with those presented in [14]. However, it should also be noted that the energy function used in these tests, equation 1, is not related to the lattice-based energy function that was used in the “competition”. Hence, one should not expect the lattice-based results to provide a reasonable approximation to the “relaxed” 3-dimensional folded states of a molecules when not restricted to a lattice.

5. A More Detailed Polypeptide Model

As previously stated, by using a simplified polypeptide model, the complexity of the problem formulation can be reduced to an acceptable level for optimization techniques. Unfortunately though, the simplifications made in section 2 do not provide a very realistic

Table 4.3 Comparison of Energies for Sequences #8 and #10

	Sequence #8	Sequence #10
F#8	-87.57	+39.48
F#10	-23.20	-97.22
Relaxation from F#8	-87.57	-33.90
Relaxation from F#10	-54.28	-97.22

model of true protein sequences. The simplifications were made only to illustrate and test the applicability of the CGU global optimization algorithm to protein structure prediction. Hence, for the CGU algorithm to be a practical method for determining tertiary structure, it must be applied to a more detailed and realistic polypeptide model.

In real proteins, each residue in the primary sequence is characterized by its backbone components NH-C_αH-C′O and one of 20 possible amino acid sidechains attached to the central C_α atom. A key element of this more detailed model is that each *sidechain* is classified as either hydrophobic or polar, and is represented by only a single "virtual" center of mass atom. Thus the potential energy function again involves only three terms: excluded volume repulsive forces between all pairs of atoms, a very powerful attractive force between each pair of hydrophobic sidechain center of mass atoms, and a torsional penalty disallowing conformations which do not exist. Since the residues in this model come in only two forms, H (hydrophobic) and P (polar), where the H-type monomers exhibit a strong pairwise attraction, the lowest free energy state is obtained by those conformations with the greatest number of H-H "contacts" (see [1], [12]). One significant advantage of this detailed formulation of the folding problem is that it allows the model to take advantage of known scientific knowledge about the chemical structure of real sequences of molecules. The use of knowledge such as the Ramachandran plot (see [4]), which specifies the allowable angles between consecutive amino acids in proteins, also greatly simplifies the problem.

Realistic molecular structure information is often given in terms of internal molecular coordinates which consist of bond lengths l (defined by every pair of consecutive backbone atoms), bond angles θ (defined by every three consecutive backbone atoms), and the backbone dihedral angles φ, ψ, and ω, where φ gives the position of C′ relative to the previous three consecutive backbone atoms C′-N-C_α, ψ gives the position of N relative to the previous three consecutive backbone atoms N-C_α-C′, and ω gives the position of C_α relative to the previous three consecutive backbone atoms C_α-C′-N. Hence, the backbone of a protein consisting of n amino acid residues can be completely represented in 3-dimensional space using these parameters, as shown in Figure 5.1.

Fortunately, these 9n-6 parameters (for an n-residue structure) do not all vary independently. In fact, some of these (7n-4 of them, to be precise) are regarded as fixed since they are found to vary within only a very small neighborhood of an experimentally determined value. Among these are the 3n-1 backbone bond lengths l between the pairs of consecutive atoms N-C′ (fixed at 1.32 Å), C′-C_α (fixed at 1.53 Å), and C_α-N (fixed at 1.47 Å). Also, the 3n-2 backbone bond angles θ defined by N-C_α-C′ (110°), C_α-C′–N (114°), and C′-N-

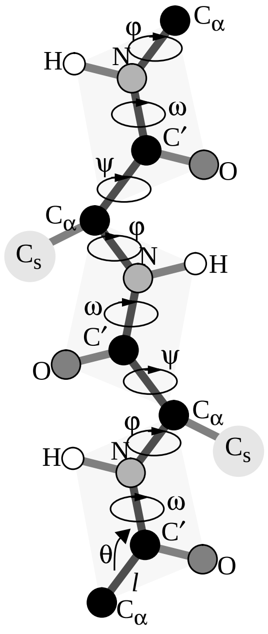

Figure 5.1 More Detailed Polypeptide Model

C_α (123˚) are also fixed at their ideal values. It is for these reasons that l and θ were also fixed in the simplified model in section 2. Finally, the n-1 peptide bond dihedral angles ω are fixed in the trans (180˚) conformation. This leaves only the n-1 backbone dihedral angle pairs (φ,ψ) in the reduced representation model. These also are not completely independent; in fact, they are severely constrained by known chemical data (the Ramachandran plot) for each of the 20 amino acid residues.

Furthermore, since the atoms from one C_α to the next C_α along the backbone can be grouped into rigid *planar* peptide units, there are no extra parameters required to express the 3-dimensional position of the attached O and H peptide atoms. These bond lengths and bond angles are also known and fixed at 1.24 Å and 121˚ for O, and 1.0 Å and 123˚ for H. Likewise, since each sidechain is represented by only a single center of mass “virtual atom” C_s, no extra parameters are needed to define the position of each sidechain with respect to the backbone mainchain. The following table (Table 5.1) of sidechain bond lengths (between the backbone atom C_α and the sidechain center of mass atom C_s), sidechain bond angles (formed by the sequence N-$C_\alpha C_s$), and sidechain torsion angles (between C_s and the plane formed by the backbone sequence N-C_α-C′) were used to fix the position of each sidechain atom. The twenty amino acids are also classified into two groups (shown in the table), hydrophobic and polar, according to the scale given by Miyazawa and Jernigan in [7].

Corresponding to this new more detailed polypeptide model is a new potential energy function. As in the simplified model of section 2, this function includes the three components in equation 1: a contact energy term E_{hp} favoring pairwise H-H residues, a steric repulsive term E_{ex} which rejects any conformation that would permit unreasonably small interatomic distances, and a main chain torsional term $E_{\varphi\psi}$ (replacing E_ξ in the simplified

Table 5.1 Residue Parameters for Center of Mass Sidechain Virtual Atoms

Residue Name	Sidechain Bond Length (Angstroms)	Bond Angle (degrees)	Torsion Angle (degrees)	H-P Designation
ALA	1.531	109.625	238.776	H
ARG	4.180	110.156	219.279	P
ASN	2.485	111.156	222.437	P
ASP	2.482	111.160	223.911	P
CYS	2.065	106.938	227.905	H
GLN	3.130	108.423	219.363	P
GLU	3.106	108.577	222.055	P
GLY	0.000	0.000	0.000	P
HIS	3.176	105.977	223.334	P
ILE	2.324	109.945	227.774	H
LEU	2.590	112.273	219.236	H
LYS	3.474	112.711	218.817	P
MET	2.976	113.370	218.790	H
PHE	3.399	112.055	222.650	H
PRO	1.868	64.159	241.896	P
SER	1.897	108.155	237.853	P
THR	2.107	109.617	231.888	P
TRP	3.907	112.930	226.091	H
TYR	3.794	109.695	222.119	H
VAL	1.968	111.792	232.308	H

model) that allows only those (φ,ψ) pairs which are permitted by the Ramachandran plot. The specific potential function used in this more detailed and accurate polypeptide model is most similar to the Sun/Thomas/Dill [13] potential function, which, as stated earlier, has already been proven successful in studies conducted independently by Sun, Thomas, and Dill and by Srinivasan and Rose [11]. In particular, the excluded volume energy term E_{ex} and the hydrophobic interaction energy term E_{hp} are defined in this case as follows:

$$E_{ex} = \sum_{ij} \frac{C_1}{1.0 + exp((d_{ij} - d_{eff})/d_w)} \text{, and}$$

$$E_{hp} = \sum_{|i-j|>2} \varepsilon_{ij} f(d_{ij}) \text{ where } f(d_{ij}) = \frac{C_2}{1.0 + exp((d_{ij} - d_0)/d_t)} .$$

The excluded volume term E_{ex} is a soft sigmoidal potential (see Figure 5.2) where d_{ij} is the interatomic distance between two C_α atoms or between two sidechain center of mass atoms C_s, d_w = 0.1 Åwhich determines the rate of decrease of E_{ex}, d_{eff} = 3.6 Å for C_α atoms and 3.2 Å for the sidechain centroids which determine the midpoint of the function (i.e. where the function equals 1/2 of its maximum value). The constant multiplier C_1 was set to 5.0 which determines the hardness of the sphere in the excluded volume interaction. Similarly, the hydrophobic interaction energy term E_{hp} is a short ranged soft sigmoidal potential (see Figure 5.3) where d_{ij} represents the interatomic distance between two sidechain centroids C_s, d_0 = 6.5 Å and d_t = 2.5 Å which represent the rate of decrease and

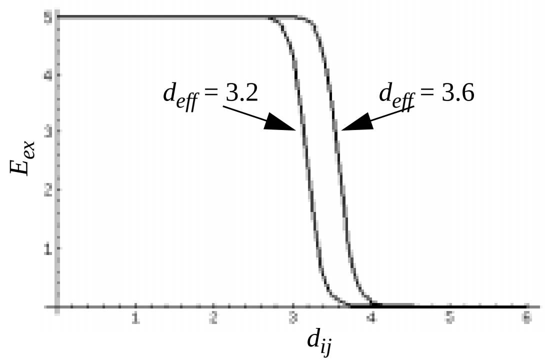

Figure 5.2 Excluded Volume Interaction Potential Function Term

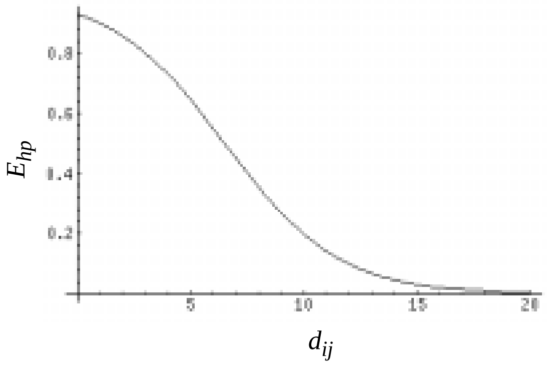

Figure 5.3 Hydrophobic Interaction Potential Function Term

the midpoint of E_{hp}, respectively. The hydrophobic interaction coefficient ε_{ij} = -1.0 when both residues *i* and *j* are hydrophobic, and is set to 0 otherwise. The constant multiplier C_2 = 1.0 determines the interaction value and is the equivalent of 1/5 of one excluded volume violation. The model is not very sensitive to the pair of constants C_1 and C_2 provided that C_1 is considerably larger than C_2.

The final term in the potential energy function, $E_{\varphi\psi}$, is the torsional penalty term allowing only “realistic” (φ,ψ) pairs in each conformation. That is, since φ and ψ refer to rotations of two rigid peptide units around the same C_α atom (see Figure 5.1), most combinations produce steric collisions either between atoms in different peptide groups or between a peptide unit and the side chain attached to C_α (except for glycine). Hence, only certain specific combinations of (φ,ψ) pairs are actually observed in practice, and are often conveyed via the Ramachandran plot, such as the one for threonine (THR) in Figure 5.4, and the φ-ψ search space is therefore very much restricted. The energy term $E_{\varphi\psi}$ accounts for this.

To compare the simplified and more detailed models at this point, it should be clear that the simplified “string of beads” model treated both φ and ψ *together* as a single “virtual dihedral angle” (denoted by ξ), thereby reducing the number of independently varying parameters from 2*n*-2 to only *n*-1 (compare Figure 5.1 with Figure 2.1). In the simplified model, each of the backbone components NH-C_αH-C′O and the associated sidechain mol-

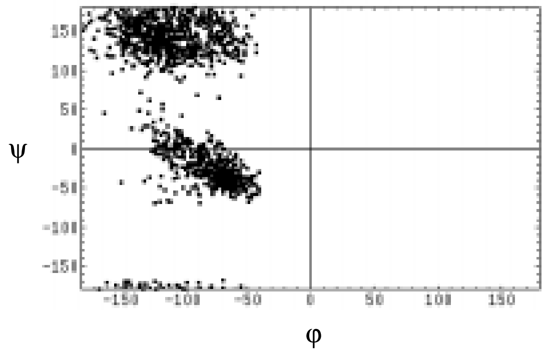

Figure 5.4 Ramachandran Plot for Threonine (THR)

ecules $C_\alpha C_s$ were replaced by a *single* backbone "virtual atom" denoted in Figure 2.1 by C (note that the backbone bond angles θ and backbone bond lengths *l* were fixed at their ideal values, even though they *do not* represent the same quantities as shown in Figure 5.1). In addition, because the single dihedral angle ξ effectively replaced the (φ,ψ) pair for each residue, there were no corresponding peptide planes and no sidechain molecules in the simplified model. However, in order to retain the H-P effects that are encoded within the backbone sequence, each "virtual atom" C in that model was categorized as either H or P, thus incorporating the hydrophobic nature of what would have been the associated sidechain (if it were represented). Hence, the CGU algorithm can be applied *unchanged* to this new more realistic model using the differentiable potential energy function $E_{total}(\phi)$ from equation 1 (with the new definitions for E_{ex}, E_{hp}, and $E_{\varphi\psi}$), where $\phi \in \mathbf{R}^\tau$ with $\tau=2n-2$ in place of n-1.

Computational testing of the CGU algorithm using this new detailed polypeptide model with the Sun/Thomas/Dill potential energy function on actual protein sequences is presented in the separate paper by Dill, Phillips, and Rosen [2].

6. Conclusions

Preliminary computational testing of the CGU algorithm applied to a simplified polypeptide model has demonstrated that the method is practical for both the homopolymer and heteropolymer models and for sequences with as many as 48 monomers. Furthermore, since the CGU algorithm is a global optimization method which is not model specific, it can be applied *unchanged* to the more detailed polypeptide model, or to any other protein model which depends on finding the global minimum of a differentiable potential energy function.

7. Acknowledgments

The authors wish to acknowledge Professor David Ferguson and his colleagues in the Department of Medicinal Chemistry at the University of Minnesota for their valuable contributions on the simple polypeptide model.

8. References

1. K.A. Dill, *Dominant Forces in Protein Folding*, Biochemistry **29** (1990), 7133-7155.
2. K.A. Dill, A.T. Phillips, and J.B. Rosen, *Molecular Structure Prediction by Global Optimization*,

Journal of Global Optimization, submitted (1996).

3. D.A. Hinds, and M. Levitt, *Exploring Conformational Space with a Simple Lattice Model for Protein Structure*, Journal of Molecular Biology **243** (1994), 668-682.
4. A.L. Lehninger, *Biochemistry: The Molecular Basis of Cell Structure and Function*, Worth Publishers, New York, 1970.
5. M. Levitt, and A. Warshel, *Computer Simulation of Protein Folding*, Nature 253 (1975), 694-698.
6. C.D. Maranas, I.P. Androulakis, and C.A. Floudas, *A Deterministic Global Optimization Approach for the Protein Folding Problem*, Dimacs Series in Discrete Mathematics and Theoretical Computer Science, in press (1995).
7. S. Miyazawa, and R.L. Jernigan, *A New Substitution Matrix for Protein Sequence Searches Based on Contact Frequencies in Protein Structures*, Protein Engineering **6** (1993): 267-278.
8. A. Monge, R.A. Friesner, and B. Honig, *An Algorithm to Generate Low-Resolution Protein Tertiary Structures from Knowledge of Secondary Structure*, Proceedings of the National Academy of Sciences USA **91** (1994), 5027-5029.
9. A.T. Phillips, J.B. Rosen, and V.H. Walke, *Molecular Structure Determination by Convex Global Underestimation of Local Energy Minima*, Dimacs Series in Discrete Mathematics and Theoretical Computer Science **23** (1995), 181-198.
10. J. Skolnick, and A. Kolinski, *Simulations of the Folding of a Globular Protein*, Science **250** (1990), 1121-1125.
11. R. Srinivasan and G.D. Rose, *LINUS: A Hierarchic Procedure to Predict the Fold of a Protein*, PROTEINS: Structure, Function, and Genetics **22** (1995), 81-99.
12. S. Sun, *Reduced representation model of protein structure prediction: statistical potential and genetic algorithms*, Protein Science **2** (1993), 762-785.
13. S. Sun, P.D. Thomas, and K.A. Dill, *A Simple Protein Folding Algorithm using a Binary Code and Secondary Structure Constraints*, Protein Engineering, submitted (1995).
14. K. Yue, K.M. Fiebig, P.D. Thomas, H.S. Chan, E.I. Shakhnovich, and K.A. Dill, *A Test of Lattice Protein Folding Algorithms*, Proceedings of the National Academy of Sciences USA **92** (1995), 325-329.

K.A. Dill, Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, CA 94118

A.T. Phillips, Computer Science Department, United States Naval Academy, Annapolis, MD 21402

J.B. Rosen, Computer Science and Engineering Department, University of California San Diego, San Diego, CA 92093